Anàlisi de dades òmiques

Prova d'avaluació contínua 1

Martina Pérez Pérez

20-3-2025

Taula de continguts

Abstract / Resum

En aquest treball es presenta una anàlisi exploratòria de dades metabolòmiques orientada a la identificació de biomarcadors per al diagnòstic del càncer gàstric. Utilitzant un conjunt de dades públic del *Metabolomics Workbench*, s'ha implementat un flux d'anàlisi complet mitjançant eines bioinformàtiques modernes de l'ecosistema R/Bioconductor. Les dades, obtingudes de mostres d'orina, s'han estructurat en un objecte *SummarizedExperiment*. L'anàlisi multivariant, que inclou un Anàlisi de Components Principals (PCA) i clustering jeràrquic, ens permet reduir la dimensionalitat de les dades i identificar patrons subjacents que diferencien parcialment els grups de mostres (controls sans, pacients amb càncer gàstric i altres condicions). Els resultats mostren que les dues primeres components principals capturen aproximadament el 35% de la variància total, i l'agrupació en tres clústers revela estructures relacionades amb els grups clínics predefinits. Aquest estudi demostra la utilitat de les tècniques d'anàlisi multivariant en dades metabolòmiques complexes i estableix les bases per a futurs estudis de validació amb cohorts més àmplies.

Objectius

L'anàlisi de dades òmiques és una eina fonamental en la recerca biomèdica moderna, que ha permès explorar de manera integral els perfils moleculars i les seves alteracions en diferents condicions biològiques. Mitjançant l'ús de tècniques de metabolòmica, és possible caracteritzar el conjunt de metabòlits presents en mostres biològiques, proporcionant informació valuosa sobre els processos bioquímics subjacents. Aquest treball se centra a desenvolupar un flux d'anàlisi simplificada però completa per a dades de metabolòmica, utilitzant eines bioinformàtiques modernes com *Bioconductor* i tècniques d'anàlisi multivariada a partir de conjunts de dades complexes.

Per tant, els objectius principals són:

- Desenvolupar el procés d'anàlisi de dades òmiques utilitzant un conjunt de dades.
- Crear i manipular un objecte de classe *SummarizedExperiment* que permeti emmagatzemar i gestionar eficientment les dades metabolòmiques i les seves metadades associades.
- Realitzar un anàlisi exploratori multivariat de les dades que proporcioni una visió general dels patrons, tendències i possibles relacions entre les diferents variables metabòliques de les dades.

- Utilitzar les eines i paquets de *Bioconductor* per al processament i visualització de dades òmiques.
- Crear un repositori a GitHub i pujar tots els arxius i documents relacionats amb l'anàlisi.

Mètodes

Origen i naturalesa de les dades

Per a aquest estudi s'ha utilitzat el conjunt de dades metabolòmiques utilitzades a l'estudi "*1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer*" amb doi: 10.21228/M8B10B, obtingut del repositori públic *Metabolomics Workbench*. Aquest dataset conté perfils metabolòmics obtinguts, permetent la identificació i quantificació de diferents metabòlits de les mostres analitzades.

Preprocessament de les dades

Per preprocessar les dades, s'han organitzat i estructurat en un objecte de classe *SummarizedExperiment*, un format estàndard proporcionat per *Bioconductor* que permet emmagatzemar conjuntament les dades metabolòmiques, informació de les característiques (metabòlits) i informació de les mostres. Aquesta estructura facilita la manipulació i la integració dels diferents nivells d'informació presents a l'estudi.

Eines bioinformàtiques

Tot l'anàlisi s'ha fet utilitzant el llenguatge de programació R a RStudio amb la versió més actualitzada juntament amb diversos paquets de *Bioconductor*. També s'han utilitzat paquets i llibreries com:

- *SummarizedExperiment*: per a la gestió de les dades òmiques
- *ggplot2*: per a la visualització de dades
- *ggrepel*: per les etiquetes als gràfics
- *knitr*: per generar informes dinàmics a R Markdown

Anàlisi exploratori

L'anàlisi exploratori s'ha estructurat en quatre parts:

1. **Anàlisi descriptiva general:** S'ha realitzat una caracterització del conjunt de dades, incloent estadístiques descriptives bàsiques, exploració de valors perduts i distribució dels metabòlits. S'han fet servir representacions gràfiques com histogrames per visualitzar la distribució de les dades.

2. **Anàlisi de Components Principals (PCA):** S'ha implementat un PCA per reduir la dimensionalitat de les dades i visualitzar les principals fonts de variació. Aquest anàlisi ens permet identificar possibles agrupacions naturals entre les mostres i avaluar la contribució de cada metabòlit a la variabilitat observada al conjunt de dades.
3. **Anàlisi de clústers:** S'han aplicat tècniques de clustering jeràrquic per identificar patrons de similitud tant entre mostres com entre metabòlits. L'agrupament permet la identificació de subgrups de mostres amb perfils metabolòmics similars.

Resultats

Dades emprades

El conjunt de dades escollit representa unes dades de metabolòmica de mostres d'orina per identificar biomarcadors potencials que puguin diferenciar entre pacients amb càncer gàstric i altres grups. Un cop processades les dades, obtenim els tres conjunts següents:

1. Matriu de dades

   La matriu de dades conté 140 mostres i 149 metabòlits. Cada cel·la representa la concentració d'un metabòlit específic en una mostra determinada.
2. Metadades metabolòmiques

   Les metadades metabolòmiques inclou la informació sobre els 149 metabòlit mesurats. Aquesta matriu conté 5 variables descriptives per a cada metabòlit:
   - Idx: Identificador
   - Name: Nom del metabòlit
   - Label: Etiqueta descriptiva del metabòlit
   - Perc_missing: Percentatge de valors faltants per a cada metabòlit
   - QC_RSD: Puntuació de qualitat que representa la variació en les mesures d'aquest metabòlit en totes les mostres.
3. Metadades de mostres

   La matriu de metadades de les mostres conté informació sobre les 140 mostres analitzades amb 4 variables descriptives:
   - Idx: Identificador
   - SampleID: Identificador
   - SampleType: Indica si la mostra és un QC agrupat o una mostra d'estudi.
   - Class: Indica el resultat clínic observat per a aquest individu.

A partir dels tres conjunts de dades mencionats anteriorment, hem creat l'objecte *SummarizedExperiment*.

SummarizedExperiment i ExpressionSet

L'anàlisi de dades òmiques requereix estructures de dades que permetin emmagatzemar i gestionar tant els mesuraments experimentals com la informació associada a mostres i característiques. Una de les principals diferències entre *SummarizedExperiment* i *ExpressionSet* és la matriu d'assaig. L'objecte *SummarizedExperiment* pot contenir diverses matrius de dades dins d'un mateix objecte mitjançant la llista *assays*, mentre que *ExpressionSet* només permet una única matriu. *ExpressionSet* està més orientat a experiments d'expressió gènica, mentre que *SummarizedExperiment* és més general i pot utilitzar-se per a qualsevol tipus de dades experimentals. Les metadades a *SummarizedExperiment* estan estructurades perquè té un sistema més robust. Una altra diferència interessant és que *SummarizedExperiment* es manté dins de *Bioconductor* i s'integra millor amb els paquets d'anàlisi moderns.

De totes maneres, *SummarizedExperiment* manté compatibilitat amb *ExpressionSet*, i això facilita la migració del codi.

Anàlisi de Components Principals

L'Anàlisi de Components Principals (PCA) s'ha realitzat amb l'objectiu de reduir la dimensionalitat dels dades metabòlomiques i visualitzar les principals fonts de variació en el conjunt de mostres.

PC	Variancia	Acumulada
1	26.793333	26.79333
2	7.978222	34.77156
3	5.502297	40.27385
4	4.665436	44.93929
5	3.713754	48.65304
6	3.473927	52.12697
7	3.039285	55.16625
8	2.515232	57.68149
9	2.289617	59.97110
10	2.111303	62.08241
11	1.965153	64.04756
12	1.819321	65.86688
13	1.612687	67.47957
14	1.576777	69.05634
15	1.502776	70.55912
16	1.444358	72.00348
17	1.324648	73.32813
18	1.213375	74.54150
19	1.165081	75.70658
20	1.147772	76.85436

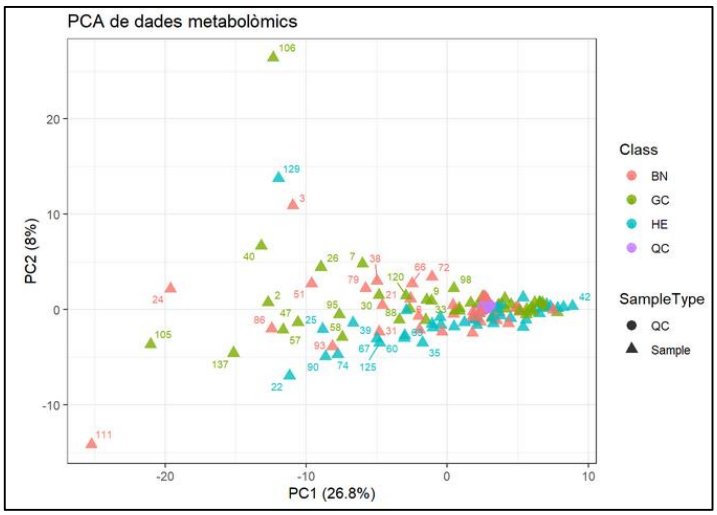
A la taula anterior podem veure un resum del percentatge de variància explicada per cada un dels primers 20 components principals, així com la variància acumulada.

La primera component principal (PC1) explica el 26,79% de la variància total, el que indica que es la direcció principal de variabilitat a les dades.

La segona component principal (PC2) explica el 7,98% addicional de la variància, elevant la variància acumulada explicada per les dues primeres components al 34,77%.

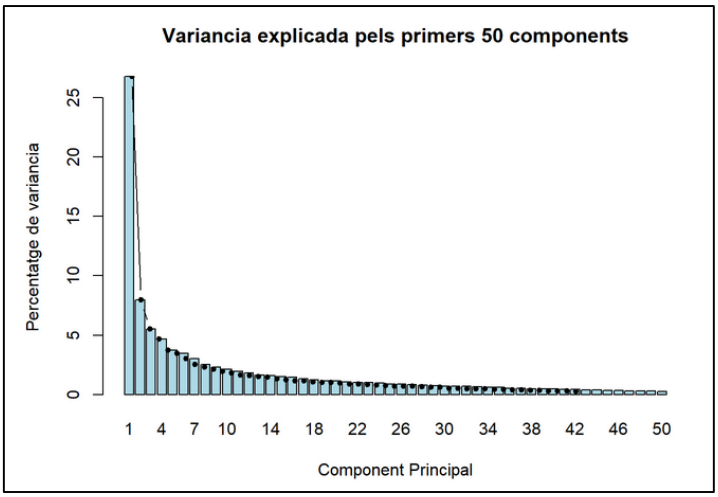
Els següents components expliquen quantitats decreixents de variància. Això suggereix que una part considerable de la variabilitat en les dades es pot representar en un espai de menor dimensió utilitzant les primeres components principals.

Els gràfics de dispersió de PC1 i PC2 permeten visualitzar si existeixen agrupacions de mostres basades en la classe o el tipus de mostra a l'espai dels dos primers components principals.



Els punts tenen un color o un altre segons la classe de la mostra i la forma del punt indica el tipus de mostra.

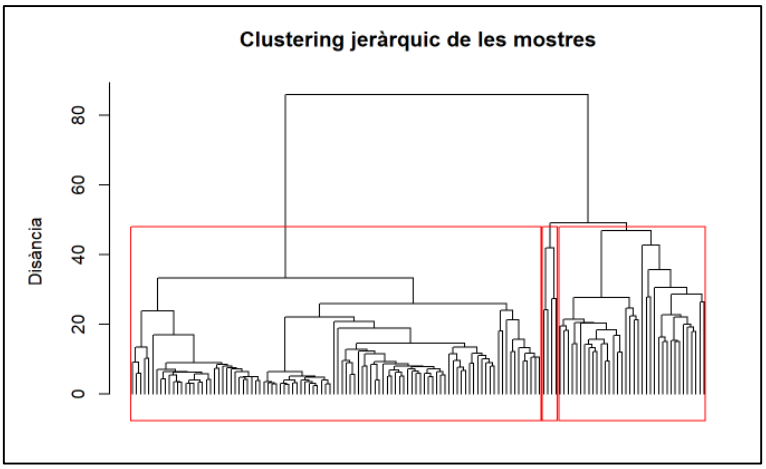
A més, el gràfic de barres (*scree plot*) mostra el percentatge de variància explicada per les primeres 50 components principals.



Aquest gràfic determina el nombre de components principals que retenen una quantitat significativa d'informació de les dades. La disminució a la variància explicada a mesura que augmenta el nombre de components suggereix un punt d'inflexió a partir del qual els components posteriors contribueixen poc a la variància total.

Anàlisi de clústers

L'anàlisi de clustering jeràrquic identifica els grups de mostres amb perfils metabolòmics similars. A partir de l'algorisme *hclust* amb el mètode *ward.D2* per minimitzar la variància dins de cada clúster, generem la visualització del dendrograma per visualitzar l'estructura jeràrquica dels clústers.

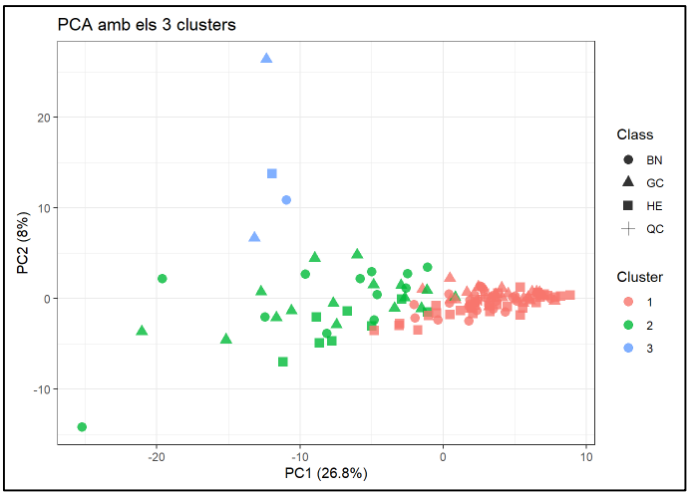


Tallant el dendrograma a 3 clústers, s'identifiquen el nombre de mostres que pertanyen a cadascun. La taula següent mostra la distribució:

clusters	Freq
1	100
2	36
3	4

Això indica que el clúster 1 conté 100 mostres, el clúster 2 conté 36 mostres i el clúster 3 conté 4 mostres.

La superposició de la informació dels clústers al gràfic de dispersió de PC1 contra PC2 permet observar la distribució dels clústers a l'espai de variabilitat principal de les dades metabolòmiques.



El clúster 1 és més heterogeni quant a la representació de les classes. El clúster 2 té més mostres de les classes BN i GC, i tendeix a ocupar una regió específica a l'espai PCA, encara que amb certa superposició amb altres clústers. De manera semblant, el clúster 3, més petit, es localitza més amunt que la resta.

Discussió

S'ha realitzat una anàlisi exploratòria de dades metabolòmiques urinàries per al diagnòstic del càncer gàstric. A través de l'Anàlisi de Components Principals (PCA) i l'anàlisi de clustering jeràrquic, s'han identificat patrons en les dades que suggereixen l'existència d'una estructura subjacent relacionada, encara que no exclusivament, amb les classes de mostra predefinides (BN, GC, HE, QC).

És important reflexionar sobre les limitacions d'aquest estudi. En primer lloc, la mida de la mostra (140 mostres), podria beneficiar-se d'un nombre més gran de participants per assegurar la robustesa i la generalització dels resultats. A més, la distribució de les mostres entre les diferents classes no és homogènia, cosa que podria influir en la capacitat per detectar diferències subtils entre els grups.

En segon lloc, la presència de valors faltants requereix la imputació utilitzant la mediana. Si bé aquest és un mètode comú, és una simplificació que podria no reflectir la veritable distribució de les dades faltants i fins i tot introduir cert biaix als anàlisis posteriors.

Com a treball futur, seria crucial validar aquests resultats en una cohort independent i més gran. S'haurien d'explorar mètodes estadístics més avançats per identificar els metabòlits que contribueixen de manera més significativa a la separació entre les classes clíniques o els clústers identificats. Finalment, es podria considerar la integració d'aquestes dades metabolòmiques

amb altres fonts d'informació clínica i òmica per obtenir una comprensió més completa del problema biològic.

Conclusions

L'anàlisi de dades metabolòmiques urinàries dut a terme en aquest treball ha permès explorar el potencial d'aquestes dades en la identificació de patrons associats al càncer gàstric. Les principals conclusions són:

1. L'ús d'estructures de dades especialitzades com *SummarizedExperiment* facilita la manipulació i integració de dades òmiques complexes, proporcionant un marc robust per a l'anàlisi bioinformàtic.
2. L'Anàlisi de Components Principals ha revelat que aproximadament el 35% de la variància a les dades pot ser explicada només amb dues components, indicant l'existència de patrons destacables en els perfils metabolòmics.
3. El clustering jeràrquic ha identificat tres grups principals de mostres.
4. La distribució de les mostres en l'espai de les components principals suggereix l'existència de diferències metabolòmiques entre els grups clínics, especialment entre pacients amb càncer gàstric i controls sans.
5. Les limitacions identificades, com la mida mostral i la presència de valors faltants, assenyalen la necessitat de validar aquests resultats en cohorts independents i més grans.

En conjunt, en aquesta pràctica hem pogut veure la viabilitat d'aplicar tècniques d'anàlisi multivariant per explorar dades metabolòmiques i identificar amb precisió biomarcadors específics del càncer gàstric en mostres d'orina.

Referències

1. *Metabolomics Workbench : NIH Data Repository*. https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&DataMode=NMRData&StudyID=ST001047&StudyType=NMR&ResultType=1#DataTabs
2. Loriot, L. G. A. A. (2023, 5 octubre). *Chapter 2 SummarizedExperiments | RNA-Seq analysis with R and Bioconductor*. https://uclouvain-cbio.github.io/bioinfo-training-02-rnaseq/sec-se.html
3. Sanchez, A. *Introduction to microarray data exploration and analysis with basic R functions*. https://aspteaching.github.io/Analisis_de_datos_omicos-Ejemplo_0-Microarrays/ExploreArrays.html#31_Univariate_statistical_analysis

4. *Tutorial1*. https://cimcb.github.io/MetabWorkflowTutorial/Tutorial1.html
5. El codi i les dades es troben a un repositori de GitHub i es pot accedir amb aquest enllaç: https://github.com/martinaperezp/Perez-Perez-Martina-PAC1#